SharedCon: Implicit Hate Speech Detection using Shared Semantics

Anonymous ACL submission

Abstract

The ever-growing presence of hate speech on social network services and other online platforms not only fuels online harassment, but also presents a growing challenge for hate speech detection. As this task is akin to binary classification, one of the promising approaches for hate speech detection is the utilization of Contrastive Learning (CL). However, recent studies show that CL, on its own, may not be sufficient for the nuanced task of detecting implicit hate speech - the struggle of which can be attributed to the subtle nature and context dependency of such pejorative remarks. Thus, researchers proposed a modified CL approach equipped with extra modules such as human-written implications or machine-generated augmented data for better implicit hate speech detection. While this approach can potentially enhance the overall performance due to its additional data general, it runs the risk of overfitting as well as heightened cost and time to obtain. These drawbacks serve as motivation for us to design a methodology that is not dependent on human-written or machine-generated augmented data for training. Thus, we propose a straightforward, yet effective, clustering-based contrastive learning approach that leverages the shared semantics among the data.

1 Introduction

Warning: *this paper contains content that may be offensive or upsetting.*

Detecting hate speech is more challenging compared to other tasks due to its varying perception and definition among individuals, which is reflected in the inconsistencies between datasets (Luo et al., 2023). Implicit hate speech, in particular, often lacks explicit cues such as swear words or insults, making its detection even more challenging. In the realm of implicit hate speech detection, the SOTA performance has been achieved through contrastive learning. Recently, researchers suggested a few models for identifying implicit hate speech (Sridhar and Yang, 2022; Lin, 2022). While these approaches exhibit promising performance for the in-dataset scenario, the performance is significantly degraded for the cross-dataset scenario (Wiegand et al., 2019) as a result of their inconsistent nature. Previous research has explored methods such as data augmentation (DA) to generate positive pairs, or creating implicit meanings (*implication*) of hateful posts manually by human thus establishing them as positive pairs. Kim et al. (2022) claimed the importance of model's generalization at implicit hate speech detection while simultaneously comparing the two, and proposed a contrastive learning approach by utilizing a common implication in certain datasets. They introduced a robust training strategy that exploits common implications for training hate speech detection models.

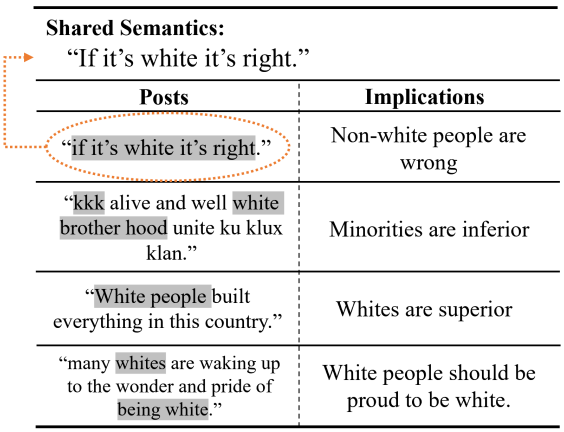


Figure 1: An example of shared semantics among posts. We often find a representative post among the posts sharing similar semantics, which is faster and even more accurate than human-written implications. The highlighted portion represents the common meaning shared among the posts.

These approaches show good performance for in-domain performance, yet such extra modules make

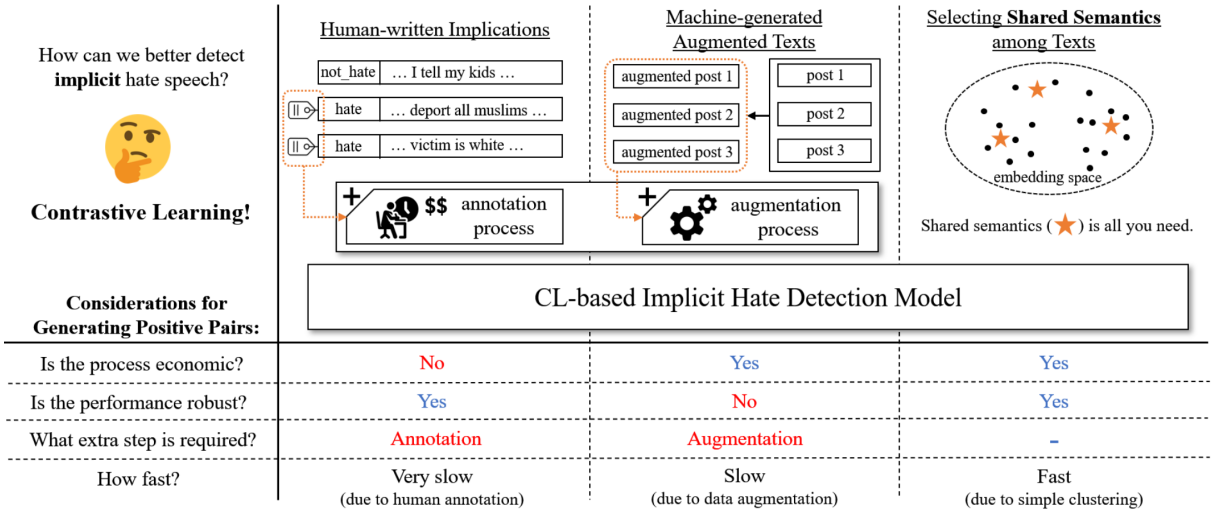


Figure 2: **Middle**: The DA approach does not require special annotations but suffers from lower performance. **Left**: the *implication* approach yields better performances, but its reliance on implication-labeled datasets presents drawbacks in terms of cost and time for creation. **Right**: we propose a novel methodology called *SharedCon* that effectively detects implicit hate speech without directly relying on human-generated implications, thereby addressing economic concerns and enhancing the robustness of implicit hate speech detection.

these approaches work well on the out-domain case; in other words, they are hard to generalize to out-domain cases. Especially, for implicit hate speech detection, we observe that it is crucial to find similar texts sharing the same implicit hate or abusiveness. However, there are some limitations to these approaches: (1) It requires experts to manually generate implications, which is expensive. For example, for an implicit hate speech dataset like IHC (ElSherief et al., 2021), total annotation cost was $15k and it took three years to complete the dataset construction. This limitation significantly restricts the scalability of this approach when applied to a wide range of datasets or different domains (Sheth et al., 2023). Furthermore, (2) It is hard to say that DA approach uses the correct positive pairs. The DA strategy merely employs post augmentation as a positive pair, which lacks semantic consistency across posts. Figure 2 shows the comparison between the existing CL appraches and our methodology. Hence, utilizing implications crafted by humans proves more effective than the DA approach, but is susceptible to errors. We conducted error analysis in Section 5.5 for human-written implications of IHC dataset.

In order to address implication costs and enhance the existing approaches, we propose ***SharedCon***, a simple yet effective technique—leveraging shared semantics from the training data itself without additional data construction step. We extract shared semantics of training datasets by grouping posts, and consider the center of each group as the representative semantics shared by the posts. *SharedCon* is evaluated on three hate speech benchmark datasets in both in-dataset and cross-dataset settings. Experimental results demonstrate that *SharedCon* outperforms our baselines with an 0.43% average improvement in the in-dataset setting and 1.43% average improvement in the cross-dataset setting. Furthermore, we compare the performance of *SharedCon* with current SOTA model (*ImpCon*). The results show that *SharedCon* can deliver comparative (even outperform for some cases) performance, against the current SOTA model that requires a human-labor-intensive step. Our empirical analysis displays that utilizing shared semantics in the quality-assured sentence embedding space can work as effective as manually produced implications. While the representation of shared semantics works similarly to the implication in Kim et al. (2022), our proposed process provides the flexibility to train the model on various datasets, even those without human-annotated implications. The proposed approach is simple yet effective without extra modules. We cluster the training data according to semantics and select one sample as a core shared semantic sample. Then, the selected sample plays a role of the anchor in the CL step.

Our main contributions are as follows:

- We propose a straightforward and effective technique—leveraging shared semantics from

the training data without the need for human-annotated implications.

- We verify the efficacy of shared semantics by enhancing performance across both in-dataset and cross-dataset scenarios.
- We achieve performance that is comparable to, and in some cases, even outperform the current SOTA model, all without the need for human labor-intensive steps.

2 Related Work

Hate Speech Detection There are several datasets on explicit hate speech with different purposes (Warner and Hirschberg, 2012; Davidson et al., 2017; Founta et al., 2018; Basile et al., 2019). Davidson et al. (2017) addressed distinguishing hate speech from offensive language by using a crowd-sourced hate speech lexicon. Implicit hate speech, on the other hand, characterized by its subtle linguistic expressions, presents a more complex obstacle for detection compared to its explicit hate speech (Ocampo et al., 2023). Recently, researchers have started considering implicit hatred or abusiveness. Han and Tsvetkov (2020) introduced a novel generic framework, MACAS, for online abusive language detection across different domains. Wiegand et al. (2021) addressed the challenges in detecting implicitly abusive language, highlighting the limitations of existing datasets and proposed a divide-and-conquer strategy to categorize subtypes of implicit abuse. In order to identify implicit hate speech, implicit data is necessary. ElSherief et al. (2021) provided a benchmark corpus to understand implicit hate speech beyond explicit forms. Leveraging the implicit hate speech dataset, Kim et al. (2022) introduced *ImpCon*, a contrastive learning method that addresses the cross-dataset evaluation challenge in implicit hate speech detection.

Contrastive learning Recently, Contrastive learning has become a prominent tool in unsupervised representation learning in various domains. SimCLR (Chen et al., 2020) randomly selects two augmented examples from a minibatch as a positive pair, while the other augmented examples within the minibatch are treated as negative pairs. Khosla et al. (2020) proposed to select positive pairs and negative pairs by considering anchor points. In the natural language processing field, SimCSE (Gao et al., 2021) selected positive pairs by using annotated pairs from NLI datasets, where entailment pairs are considered as positives and contradiction pairs used as hard negatives. Gunel et al. (2021) suggested using contrastive learning as a fine-tuning tool to capture similarities between examples from same classes and contrast them with examples from other classes to improve generalization. LCL (Suresh and Ong, 2021) performed well on fine-grained text classification tasks by adaptively weighting the relationships between classes to distinguish between more difficult samples. We also used contrastive learning to fine-tune model performance.

Clustering for contrastive learning Yang et al. (2023) performs cluster-level contrastive learning to incorporate measurable emotion prototypes. Gao et al. (2022) proposed weakly supervised contrastive learning method that allows us to consider multiple positives and multiple negatives. This is prototype-based clustering method that avoids semantically related events being pulled apart. We suggest selecting multiple positives by clustering to use contrastive learning in the fine-tuning process for shared semantics without human-written implications.

3 Contrastive Learning with Shared Semantics

While the idea of utilizing shared semantics in the fine-tuning process is promising for implicit hate speech detection, still one major drawback occurs; it requires a dataset with implications, which are written by humans manually, for implicit hate speech detection. Thus, we study an effective way of finding shared semantics without human involvement and improving the cross-dataset performance.

3.1 Shared Semantics

We use the term **Shared Semantics** to describe the approach for finding shared semantics without human involvement, which we presume represents the shared characteristics inherent in instances of real-world data. We provide Figure 1 to show an example of shared semantics we extracted from IHC (ElSherief et al., 2021) dataset. As shown in Figure 1, we observe that shared semantics effectively captures the latent meaning of hate speech posts as 'implication', a human-generated label representing the implicit meaning of a post. As can be seen in Figure 3, we first extract sentence

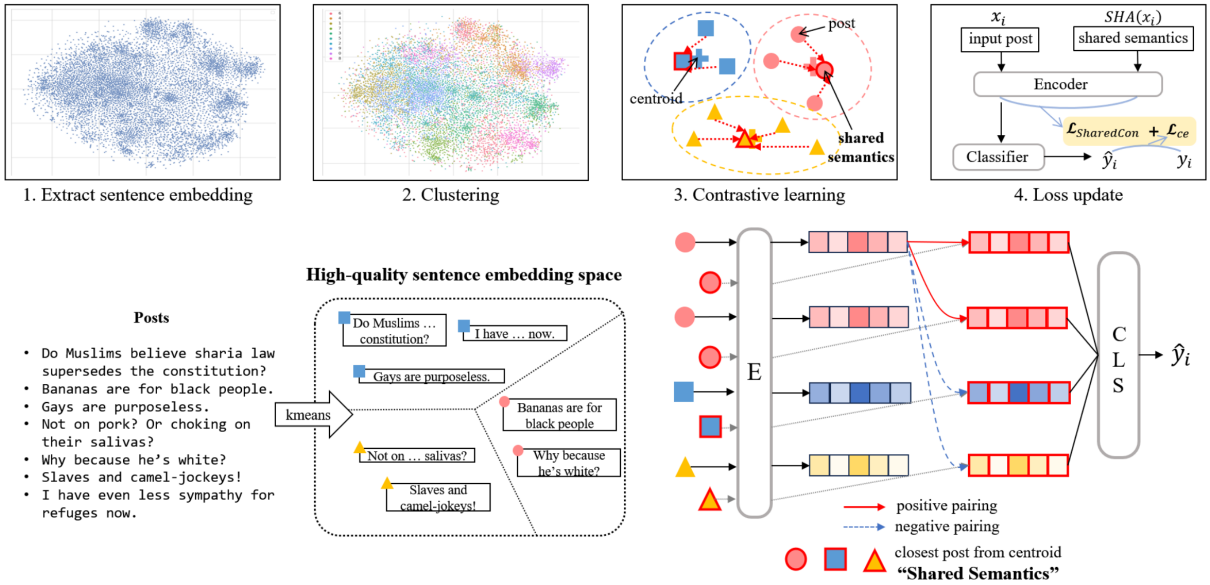


Figure 3: Overall workflow of *SharedCon*. With a centroid sentence(*i.e.,*, an anchor) from the clustered sentence embedding space, we train to make sentences within the same cluster closer and sentences from different clusters more distant. 'E' stands for an encoder, while 'CLS' signifies a classifier. Detailed information will be addressed in section 3.

embeddings of the training data using the existing sentence embedding models. Then we cluster these embeddings using K-means algorithm. Each cluster formed in this manner is expected to share a common semantic meaning. For sampling our positive set, we consider the centroid posts as the closest sentence from the corresponding cluster centers. Let $c = \{c_1, c_2, \ldots, c_K\}$ be the set of centroid posts, where c_i is the centroid post of the cluster C_i of K clusters $C = \{C_1, C_2, \ldots, C_K\}$. Given a batch size N, we denote the sample, centroid sample, and label triplets of ith input as $\{x_i, c_i, y_i\}_{i=1}^{N}$ and the positive sample of x_i as x_i^{pos} $(i \geq 1)$. The model is trained to bring samples, represented as a pair of posts and centroid posts, within the same cluster closer together and push away those from different clusters.

We will conduct a detailed analysis of the locations of shared semantics within a dataset in 5.

3.2 Training Objective

We propose SharedCon which adjusts contrastive learning (CL) loss [Eq. 1] for implicit hate speech detection. During training, the classifier learn the implication of hate speech posts through the CL loss [Eq. 2] and simultaneouly be trained binary classification of hate or not using the cross-entropy (CE) loss [Eq. 3]. With a batch of N training samples, in cases where x_i^{pos} becomes the jth input in a batch, *i.e.,*, $x_i^{pos} = x_j$, we consider $j = i + N$ if $i \leq N$, and $j = i - N$ if $i > N$. That is, given that each input sample in the batch is paired with its positive, there are a total of $2N$ samples within one batch. When training input comes in, We also accept their samples as positive while the members of the same cluster are in the same mini-batch; in other words, we permit in-batch multiple positives.

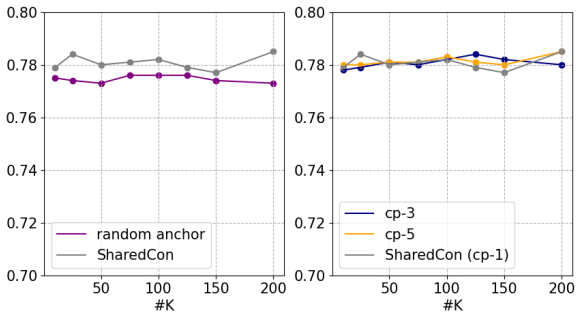


Figure 4: **Left**: The performance when anchors are randomly selected. An overall decrease in performance can be observed in random anchor selection. **Right**: The performance when samples near the centroid are selected as anchors. 'cp-3 and 5' refers to the selection of the anchor from the three and five posts most proximal to the centroid, respectively.

We use the following supervised contrastive loss $\mathcal{L}_{SCL}$ (Khosla et al., 2020) to pull multiple positive pairs close together and push negative pairs away.

$$\mathcal{L}_{SCL} = \sum_{i=1} \frac{-1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \log \frac{e^{h_i \cdot h_p / \tau}}{\sum_{k \in I \setminus \{i\}} e^{h_i \cdot h_k / \tau}}. \quad (1)$$

Motivated by Khosla et al. (2020), we modified 1 and justified *SharedCon* loss 2 and our positive set $\mathcal{P}_i$ as follow:

$$\mathcal{L}_{SharedCon} = \sum_{i=1}^{2N} \frac{-1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \log \frac{e^{h(x_i) \cdot h(SHA(x_i))/\tau}}{\sum_{k \in I \setminus \{i\}} e^{h(x_i) \cdot h(x_k)/\tau}}, \quad (2)$$

$$\mathcal{P}_i = \begin{cases} \{p \in [N+1, 2N] : (c_p = c_i) \wedge (c'_p = c'_i)\} \\ \qquad i \in [1, N] \\ \{p \in [1, N] : (c_p = c_i) \wedge (c'_p = c'_i)\} \\ \qquad i \in [N+1, 2N] \end{cases}$$

where $h(x_i)$ is the encoder output of input x_i and τ is a temperature parameter. Suppose a module $SHA(\cdot)$ constructs our positive set by assigning centroid posts. Then, we can arrange the positive sample for x_i as $x_i^{pos} = x_j = SHA(x_i)$. For $i \leq N$, $SHA(x_i)$ is the centroid post of the cluster to which x_i belongs. For $i > N$, $SHA(x_i)$ is the original input post of x_i. c'_i is the label of cluster (*i.e.*,, cluster number, which is different from number of the cluster) of the corresponding centroid post c_i.

The CE loss is a typical loss function used in hate speech detection and is defined as follows:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)], \quad (3)$$

	IHC	SBIC	DYNA
CE	77.7	83.8†	78.8†
+ *AugCon*	77.4	83.3	77.6
+ *ImpCon*	78.0†	83.6	-
+ *SharedCon*	**78.5**	**84.3**	**79.1**

Table 1: In-dataset performance in macro-F1. We report an average of three runs with different random seeds. † represents the previous SOTA score, while our score surpassing it is **bolded**.

where $\hat{y}_i$ is the model prediction for the x_i, and y_i is the ground truth.

The overall objective is:

$$\mathcal{L}_{overall} = \lambda \mathcal{L}_{ce} + (1 - \lambda) \mathcal{L}_{SharedCon}, \quad (4)$$

where λ is a scaling parameter which controls the balance between the CE loss and the CL loss. Our positive set $\mathcal{P}_i$ serves the role of allowing members within the same cluster to function as positive samples.

4 Experimental Results

4.1 Datasets

We conducted multiple experiments utilizing three different hate speech datasets. Note that IHC and SBIC have implied statements but DynaHate does not.

- **IHC** (ElSherief et al., 2021): a hate speech dataset derived from hate communities on Twitter and their followers, containing information about the target and the implied meaning. The dataset contains 6,346 implicit hate speech out of 22,056 tweets in the United States.
- **SBIC** (Sap et al., 2020): an offensive language dataset that has hierarchical annotations related to social bias, encompassing aspects such as offensiveness, target, and implied statements.
- **DynaHate** (Vidgen et al., 2021): a hate speech dataset obtained using a process that involves a combination of human input and model manipulation to deceive the model.

4.2 Baselines

Motivated by (Kim et al., 2022) we compare three baseline.

CE we use the cross entropy loss (CE) to fine-tune the model, which is a fundamental method for classifying hate speech.

AugCon utilizes *augmented inputs* as positive samples for inputs in the process of contrastive learning. While *AugCon* can be done without any additional need for human annotations (except for inputs and their labels), *AugCon* showed improved performance on some limited settings.

ImpCon utilizes *shared implications* as positive samples for hate speeches in the process of contrastive learning. While *ImpCon* offered consistently superior performance compared to other

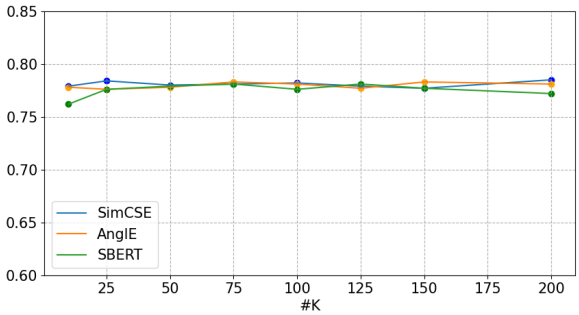


Figure 5: We compare F1 score of three different sentence embedding models. While the range of K is from 10 to 200, the performance using the three models is similar, but a slight advantage can be observed in SimCSE and AnlgE compared to SBERT. The consistency in the performance of these three models demonstrates the embedding invariance of our approach, indicating that *SharedCon* can achieve fair performance regardless of which model is used.

methods, it necessitates additional human annotation for the implication.

We design a contrastive loss that has strengths of both approaches; given that utilizing shared semantics among inputs was effective, we aim to utilize shared semantics for achieving generalization ability (= *ImpCon* strength) without additional human involvements (= *AugCon* strength).

4.3 Detection Performance

We explore the effectiveness of our method through evaluations conducted on both in-dataset and cross-dataset scenarios. We train BERT (Devlin et al., 2019) on each of the three datasets and evaluate them on those three respectively. Table 1 demonstrate the outcomes of in-dataset evaluations, with respect to sentence embedding models: SimCSE (Gao et al., 2021). Table 2 presents the evaluation results for cross-dataset scenarios, utilizing SimCSE embeddings as well.

The noteworthy thing is that the DynaHate dataset does not have implications; our approach selects appropriate representative of shared semantics from the training dataset and improves the performance. In experiments conducted in-dataset, we observe a slight improvement of approximately 0.43% over the baselines on average. However, in the cross-dataset setting, particularly for models trained on IHC, there was a significant increase of up to 4.5% in SBIC evaluation scenario.

This suggests the effectiveness of our approach, even when compared to the utilization of human-annotated implications. Overall, the experimental results show that *SharedCon* shows competitive or superior performance compared to the existing SOTA without using any human-written implications.

5 Analysis

Our methodology for defining an anchor, the sentence closest to the centroid of a cluster, is twofold: 1) in Section 5.1, we randomly select a post and allocate is as an anchor; and 2) in Section 5.2, we choose one sentence among the top three or five sentences that are most proximal (in terms of Euclidean disance) to the centroid of the cluster as the anchor.

5.1 Random Anchors

We investigated where shared semantics manifest in the embedding space by altering the method of selecting anchor posts. When a sentence from random position was chosen as an anchor for contrastive learning, the performance decreased compared to selecting centroid sentence as an anchor. Figure 4 illustrates the robustness of *SharedCon* compared to the random-anchor-selection approach. The gray graph represents macro-f1 score of *SharedCon* and the violet graph represents that of random selection. As the value of K changes, both selection strategies show minor fluctuations in macro-F1 score, but *SharedCon* consistently achieves good performance. From these experimental results, we can infer that shared semantics are located at the center of clusters.

5.2 Centroid-proximate Anchors

In order to further investigate the methodology for extracting the representative shared semantics, we additionally experiment while introducing variations to the positive pairs. Figure 4 showing two results of changing the anchor for contrastive learning from the centroid post among top 3 and 5 closest posts to the centroid. In the embedding space of SimCSE, we trained the model with IHC and selected x_i^{pos} for x_i among the three closest posts to the centroid post of the cluster. We conducted a similar experiment, changing the selection from the three closest posts to the five closest posts. For K = 200 (to compare with Table 2), both performances of 3-anchors' and 5-anchors' are comparable to that of *SharedCon*, or occasionally surpass it. This suggests that choosing centroid-around posts might play a role as noise, contributing to

Test	Train											
	IHC				SBIC				DYNA			
	CE	*Aug*	*Imp*	*Ours*	CE	*Aug*	*Imp*	*Ours*	CE	*Aug*	*Imp*	*Ours*
IHC	-	-	-	-	59.6	59.7	61.4†	**62.4**	66.0†	65.6	-	**67.0**
SBIC	56.8	58.1	60.7†	**65.2**	-	-	-	-	68.2†	66.3	-	67.9
DYNA	53.1	54.6	57.9†	**59.5**	60.3	61.2†	61.2†	**62.0**	-	-	-	-

Table 2: Performance evaluation in macro-F1 for three cross-datasets compared with *AugCon* and *ImpCon*. We report the average of three runs with different random seeds. Note that the datasets listed in the columns were used for training, while those in the rows were used for evaluation. † represents the previous SOTA score, while our score surpassing it is **bolded**.

the effectiveness of adversarial training. We hypothesize the presence of shared semantics in the center of clusters, but consider the exact location to be an open question for precise positioning. We expect the exploration of alternative methods for selecting representative points beyond the cluster centroids will be reserved for future work. This aspect remains an avenue for further investigation and enhancement in our research. Hence, it can be observed that the centroid sentence performs most effectively in embodying the role of shared semantics.

5.3 Generalizability

While our model slightly outperforms in-dataset results compared to our baselines, as Kim et al. (2022) tackled, the evaluation of cross-dataset is important to show the generalization ability of a model. Therefore, we conduct a cross-dataset evaluation for the models trained on the three hate speech benchmark datasets: IHC, SBIC, DynaHate (ElSherief et al., 2021; Kim et al., 2023; Vidgen et al., 2021) to check generalizability of our proposed model. Table 2 compares cross-dataset evaluation results of *AugCon*, *ImpCon*, and *SharedCon* in SimCSE (Gao et al., 2021) embedding in cross-dataset settings. We observe that IHC-trained model achieved the highest performance while evaluated in SBIC dataset. We speculate that the reason for the high performance could be attributed to the potential presence of label errors within the dataset. After suspecting potential errors within the IHC dataset, we identified certain noisy implicit labels. These labels are contributing to confusion in the classifier, resulting in a decrease in its performance. In other words, it is better to find shared semantics among the given training dataset directly instead of using human-written implications or machine-generated augmented texts.

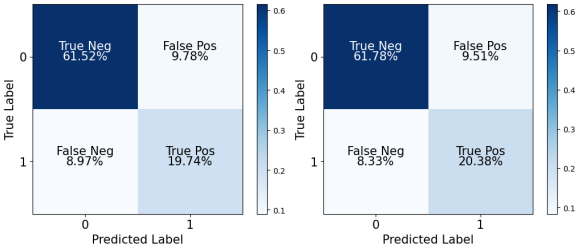
5.4 Representation by Sentence Embedding Models

We assume that the space of the sentence embedding model already effectively reflects semantic similarity. In order to validate our assumption, we also implemented an extra experiment by changing the sentence embedding model. Figure 5 shows that *SharedCon* consistently demonstrates strong performance regardless of variations in the sentence embedding model.

Sentence embedding model We leverage three different sentence embedding models: SBERT (Reimers and Gurevych, 2019), SimCSE (Gao et al., 2021) and AnglE (Li and Li, 2023)

- **SBERT** is a modified BERT using siamese/triplet structures, excels in semantic similarity and transfer learning.
- **SimCSE** enhanced sentence embeddings through contrastive learning, improving semantic understanding in textual tasks via unsupervised and supervised methods.
- **AnglE** addressed gradient issues in cosine-based methods across diverse text types and outperforms prior models in semantic similarity tasks.

By leveraging the premise that the space represented by the sentence embedding model embodies a comprehensive semantic structure, we discerned shared semantics within this framework. We aim to empirically substantiate the validity of high-quality representation space. The results for SBERT, SimCSE, and AnglE consistently exhibited favorable performance without significant deviations. In order to explore the significance of qualified sentence embedding, we further investigate our approach



(a) *ImpCon* (b) *SharedCon*

Figure 6: Confusion matrices for the model predictions on IHC test set. We evaluate the predictions made by two BERT models trained on IHC training set with (a) *ImpCon*, and (b) *SharedCon*.

by removing the sentence embedding model. After that, we randomly selected K samples within each class to constitute the positive set and trained the classifier using contrastive learning, as the way of *SharedCon*. According to Figure 4, without employing sentence embedding model, the performance has been significantly decreased in both in-dataset and cross-dataset results. For $K = 200$ and trained the model with IHC, the in-dataset score declined by 1.2%, and the cross-dataset scores decreased by 4.5% on average. This indicates the effectiveness of our methodology in extracting implications within a high-quality representation (embedding) space. In other words, the sentence embedding model provides a reasonably good representation space for finding shared semantics, and our approach of identifying positive pairs on this basis makes sense.

5.5 Error Analysis

We analyze the incorrect predictions of our approach and *ImpCon* for the IHC dataset. *SharedCon* had 355 false positives and 311 false negatives while *ImpCon* had 365 and 335 respectively. As depicted in the confusion matrices in Figure 6, *SharedCon* has 2 less false negatives and 2 less false positives compared to *ImpCon*. For *SharedCon*, out of a total of 666 incorrections in both false positive and false negative, we observed that there are 22 broken sentence posts for both in false positive and false negative. Detailed specifications regarding the number of samples are provided in 6. *ImpCon* is more vulnerable to broken sentences than *SharedCon*; clearly, broken sentences impact the model's prediction errors. Since broken sentences and obfusticated posts are commonly observed in recent social media platforms, effectively detecting these aspects is an important avenue for future research. Table 5 shows some examples of errors both *ImpCon* and *SharedCon* made. We provide samples of broken sentences that both methodologies struggle with at Table 5.

Regarding the misclassification by *SharedCon*, we point out errors within the dataset itself. We often observe errors in human-written implications in the dataset; Table 3 shows some examples that capture subtly incorrect implications. These examples illustrate instances where responses are skewed towards negative connotations or mislabeled altogether, irrespective of context. For a very basic analysis to identify the simple label error ratio, we randomly sampled 100 posts from the train set of IHC and found that approximately 27% of them were misclassified. Out of these misclassifications, 7% were ambiguous, and remaining 20% were incorrectly classified. We assume that these incorrect implications can cause further performance degradation.

P1	"when you drab the swamp in sacramento!"
I1	Immigrants should be deported
P2	"what do you expect? a white kid with air jordon shoes."
I2	Blacks have criminal tendencies
P3	"non-white hate crime mob violence comes to finland"
I3	Non-white people are violent.

* P is a post and I is its implication

Table 3: A few examples of wrong implications written by human annotations in the IHC dataset; such errors are often observed in the dataset.

6 Conclusions

We have enhanced the existing implicit hate speech detection model, originally reliant on human-generated implications. This improvement involves clustering sentence embeddings and utilizing the centroid sentence of each cluster as positive examples during the contrastive learning process. The experimental results have demonstrated that posts in the dataset have shared semantics, and our approach of using them as positive samples for contrastive learning works well. Particularly, our methodology has shown competitive or superior performance improvement of fine-tuned models in cross-dataset experiments. To sum up, our approach has enhanced model generalization without human-written implications, providing a better solution for implicit hate speech detection in terms of cost.

Limitations

While our approach reduces possible errors of wrong implications written by humans, it might be possible that a centroid sentence itself might be labeled wrong in the first place. However, it is less likely for a centroid sentence to have a wrong label since it has similar semantics compared with the other sentences in the same cluster. Another issue would be a selection process of shared-semantic sentences. Certainly, there could be a better way, but our approach—selecting a centroid sentence in a cluster—constantly shows good and stable performance compared with several other heuristics.

Ethical Consideration

Generalization Ability Enhancement Our work focuses on enhancing the generalization ability to train a hate speech detector that can consistently adapt to new forms of hate speech. Specifically, by training a model with improved performance, we aim to equip it with the capability to respond to emerging manifestations of hate speech even on the same dataset. This emphasis on enhanced generalization entails training a model that demonstrates superior performance when presented with the same dataset, enabling it to effectively address novel expressions of hate speech.

Annotator Exposure Mitigation In the past, state-of-the-art (SOTA) models exposed annotators to the risk of hate speech exposure during the process of adding implications as comments to secure such capabilities. However, our methodology has achieved comparable performance without such risks.

Risks and Potential Misuse The enhancement of these capabilities represents a positive development in the field of hate speech detection; however, it also inherently carries risks. If the improved generalization ability is exploited, it could be utilized to generate new forms of hate speech or conceal hateful behaviors. It is crucial to be vigilant about these potential risks and to carefully evaluate the use of the model and its outcomes.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, pages 491–500.

Jun Gao, Wei Wang, Changlong Yu, Huan Zhao, Wilfred Ng, and Ruifeng Xu. 2022. Improving event representation via simultaneous weakly supervised contrastive learning and clustering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3036–3049. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for

pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, pages 18661–18673.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679.

Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. Conprompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10964–10980.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Jessica Lin. 2022. Leveraging world knowledge in implicit hate speech detection. *arXiv preprint arXiv:2212.14100*.

Chu Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Legally enforceable hate speech detection for public forums. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2758–2772. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Paaras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. 2023. Peace: Cross-platform hate speech detection - a causality-guided framework. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 559–575.

Rohit Sridhar and Diyi Yang. 2022. Explaining toxic text via knowledge enhanced text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–826, Seattle, United States. Association for Computational Linguistics.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, page 19–26.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Trans. Affect. Comput.*, 14(4):3269–3280.

A Datasets Statistics

Dataset	Train set	Valid set	Test set
IHC	11,199	3,733	3,734
SBIC	35,504	4,673	4,698
DynaHate	33,006	4,125	4,124

Table 4: The statistical information of three datasets employed in our experiments.

B Comparative Analysis

Label: 1	**Post1**	**Post2**
ImpCon: 0 *Ours*: 1	"i happen to prefer the white race not partially covered in non-white ink."	"white power and proud ! ! !"
Label: 1	**Post1**	**Post2**
ImpCon: 0 *Ours*: 0	" to say " white lives matter " " is pure " " racist hate " " . anti-racist is a code word for #antiwhite . " "	"rt " : super cool german officer puts " " neo-nazi " " stickers in police van | more" ; ""

Table 5: Error examples from IHC train dataset. We compare prediction of *ImpCon* and *Ours*. Label 1 denotes that ground truth. We can observe both methodologies encounter difficulties with the broken sentences, while *ShareCon* is comparatively robust than *ImpCon* with grammatically correct sentences.

C Implementation Details

For the hyperparameter optimization, we consider various values as follows. We select the learning rate from {5e-6, 1e-5, 2e-5, 3e-5, 5e-5}, and the temperature parameter τ from {0.1, 0.3, 0.5}. We also specified λ as a scaling factor from {0.25, 0.5, 0.75}, and explored the number of cluster K from {10, 25, 50, 75, 100, 125, 150, 200}. The model archiving the highest validation F1 score is selected as the best-performing one. To ensure robustness, all experiments are executed using 5 distinct random seeds $(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)$, and the macro-F1 score on the test set is reported. We perform training on models for a span of 6 epochs employing an NVIDIA RTX 4090 GPU.